**CS697A – Topic in Computer Science – Machine Learning**

**Summer 2021**

**Project (15 Pts)**

Due date : July 30, 2021 at 11:00pm

Instructor : Zehra Cataltepe

**PURPOSE:**

Decision trees, neural networks and their combinations.

**WHERE TO SUBMIT ASSIGNMENTS**

Please submit through the class Blackboard site. Please zip and upload all your files using filename studentID\_project.zip. Submit a zip file of the Jupyter Python notebook, any additional files and also an excel file containing your expected grade from each question as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Q1 | Q2 | Q3 | Q4 | Q5 | Total |
| Max Pts | 3 | 3 | 3 | 3 | 3 | 15 |
| Expected Pts |  |  |  |  |  |  |

**POLICY:**

Collaboration in the form of discussions is acceptable, but you should write your

own answer/code by yourself. Cheating is highly discouraged for it could mean a zero or negative grade from the project.

If a question is not clear, please let me know (via email, during office hour or in class).

**Data & libraries to use:**

Use the diabetes\_train.csv and diabetes\_test.csv file provided from the [Pima Indians dataset](https://www.kaggle.com/uciml/pima-indians-diabetes-database/data).

use the scikit-learn library classifiers [MLP](https://scikit-learn.org/stable/modules/neural_networks_supervised.html), [decision tree](https://scikit-learn.org/stable/modules/tree.html), [random forest](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) and [HistGradientBoostingClassifier](https://scikit-learn.org/stable/modules/ensemble.html)

For each question, report in a table:

|  |  |  |
| --- | --- | --- |
| Parameters and their values | Training error | Test error |

**Questions:**

**Q1 [3pts]:** Use MLP (neural network) with 5 different sets of parameters (changing #hidden layers, #units in each layer, regularization, learning rate). Report the training and test errors and the parameters you used for each setting in a table as shown above.

**Q2 [3pts]:** Use decision tree with 5 different sets of parameters.

**Q3 [3pts]:** Use random forest with 5 different sets of parameters.

**Q4 [3pts]:** Use gradient boosting classifier with 5 different sets of parameters.

**Q5 [3pts]:** Combine the classifiers with the best test error you produced in Q1..Q4 using [VotingClassifier](https://scikit-learn.org/stable/modules/ensemble.html) and measure the training and test error for each of the following cases:

a) give equal weight to each classifier

b) give weight proportional to 1/(1+trainingerror)